Towards richer descriptions of our collection of genomes and metagenomes

Dawn Field, George Garrity, Jeremy Selengut, Peter Sterk, Tatiana Tatusova, Nicholas Thomson, Michael Ashburner, Jeffrey Boore, Guy Cochrane, James Cole, Claude de Pamphilus, Robert Edwards, Nadeem Faruque, Robert Feldman, Tanya Gray, Sarah Gurr, Dan Haft, Dave Hancock, Christiane Hertz-Fowler, Jennifer Hughes, Ian Joint, Matt Kane, Jessie Kennedy, Eugene Kolker, Nikos Kyrpides, Jim Leebens-Mack, Suzi Lewis, Natalia Maltsev, Victor Markowitz, Barbara Methe, Norman Morrison, Karen Nelson, Julian Parkhill, Susanna-Assunta Sansone, Andrew Spiers, Robert Stevens, Paul Swift, Chris Taylor, Yoshio Tateno, Adrian Tett, Sarah Turner, David Ussery, Bob Vaughan, Trish Whetzel, Gareth Wilson, and Anil Wipat

In this commentary, we advocate building a richer set of descriptions about our invaluable and exponentially growing collection of genomes and metagenomic datasets through the construction of consensus-driven data capture and exchange mechanisms.

Standardization activities must proceed within the auspices of open-access and international working bodies, and to tackle the issues surrounding the development of better descriptions of genomic investigations we have formed the Genomic Standards Consortium (GSC). Here, we introduce the "Minimum Information about a Genome Sequence" specification in the hopes of gaining wider participation in its development and discuss the resources that will be required to support it (standardization of annotations through the use of ontologies and mechanisms of metadata capture, exchange). As part of its wider goals, the GSC also strongly supports improving the 'transparency' of the information contained in existing genomic databases that contain calculated analyses and genomic annotations.

A wealth of genomic and metagenomic sequences

At the beginning of the genomic era, few could have imagined the wealth of data we would have amassed in a single decade. With over 4,000 genomes from eukaryotes, bacteria, viruses, plasmids, and organelles in reference collections in public databases, and an ever growing number of metagenomic datasets now coming online, considerations on how we ensure suitable stewardship of this data for the long-term is growing.

Our genome collection: More than the sum of its parts

The analysis of genomic information is impacting every area of the life sciences and beyond. A genome is the entire genetic complement of an individual and, as such, is an open door to understanding the molecular basis of organismal phenotype, how it evolves over time, and how we can manipulate it to provide new solutions to critical problems. These include potential cures for disease, drug therapies, industrial products, biodegradation of xenobiotic compounds, and even renewable energy sources. With improvements in the technology of sequencing, the growing interest in metagenomic approaches, and the proven power of comparative analysis of groups of related genomes, the day can be envisioned when sequencing tens to hundreds of genomes or more, as part of a single study, will be common place. Given current rates of genome sequencing, it has been estimated that we will have 1000 bacterial genomes in three years and more than 4,000 in ten years time [1].

Given the importance of our growing genome collection, the benefits of leveraging it through of comparative analyses, and its exponential rate of growth, it seems tantamount that we should strive to describe it as comprehensively as possible. There is increasing interest in doing so from the community for three main reasons [2-4]. The first is the 'grassroots' interest of a growing number of isolated researchers working to explain the features we see in genomes using comparative evo- and eco-genomic approaches [5]. The second is the growing need to supplement the content of a variety of databases with high level descriptions of genomes that allow useful grouping, sorting, and searching of the underlying data. The third is the growth in the number of genomes from environmental isolates and metagenomes – vast data sets of partial DNA fragments from environmental samples - that are being sequenced. The type of data generated by such studies will dwarf current stores of genomic information, and improved descriptions are of tantamount importance.

We have gaps in our current top-level descriptions of genomes for one overriding reason; we are learning with hindsight the quality and quantity of information that is required to make the description of each unique and useful. For example, strain names were not routinely captured in genome annotation documents prior to the sequencing of large number of genomes from the same species [6] but are now considered essential. Through empirical observations, we are expanding our view of the types of information that are of critical importance for testing particular hypotheses [7], exploring new patterns [5], or quantifying inherent sampling biases[5]. We are also being forced to re-think our concept of the minimal information required to adequately describe a 'genome' sequence, driven mainly by the appearance of metagenomic sequences. The number of habitats which have been sampled continues to grow. Without adequate description of the biological material used to generate these data (e.g. environmental conditions, sample processing steps before sequencing, type of sequencing method) the resulting data will be of greatly lessened value for researchers wishing to conduct subsequent comparative genomic studies. Finally, sequencing technology is advancing and the new family of methods currently being unleashed [8] will force the adoption of additional descriptors to distinguish between them.

Most often such metadata is found only in the primary literature (on a per genome basis) or in reference works, such as the recognized gold standard for bacteria, Bergey's Manual [9] (on a per species basis) [2-4]. The distributed and patchy nature of this information and the difficulties of curating even a few pieces of information for what are now hundreds of genomes makes the vision of a single definitive source of rich genomic descriptions highly desirable.

The need for co-ordinated efforts

Facilitating and accelerating the process of collecting more metadata would clearly reduce ongoing duplication of effort and maximize the ability to share and integrate data within this community. The obvious solution is to take a consensus approach that limits no one but works to support everyone.

The Genomic Standards Consortium

The GSC is an open-membership group which formed in late 2005 after two exploratory workshops that resulted directly from a call for action[4] that was circulated among the community to gauge interest in formalizing wider standardization activities. The GSC

community brings together (1) evolutionist, ecologists, molecular biologists, and other researchers analyzing collections of genomes, (2) those producing genomic databases, (3) those producing genomes and (4) computer scientists, ontology experts, and members of other standardization initiatives. These include leading members of the International Nucleotide Sequence Data Collaborators (INSDC) who are responsible for the Genbank/EMBL/DDBJ genomic databases. The guidance of the INSDC is critical to the success of this initiative as they are the official stewards of the public collection of genomes and also because the INSDC makes every effort to make sure its resources evolve in accordance with community needs.

Re-evaluating and extending the minimal information collected about genome sequences

We are working to formalize a set of additional core descriptors for genomes through the generation of a "Minimal Information about a Genome Sequence" (MIGS) specification. The draft MIGS checklist is available from the gensc.sf.net website, but is briefly described here. The information required to comply with it is routinely reported in primary genome publications (or is referenced therein) and needs only to be standardized and made available in electronic form to vastly improve public access to genomic metadata [4]. Since it was originally suggested, the MIGS [4] specification has been simplified and updated by the GSC through the normal iterative process of revision to contain (1) only curated information which can not be calculated from a raw genomic sequence, (2) core descriptors specific to the major taxonomic groups (eukaryotes, prokaryotes, plasmids, viruses, organelles and metagenomes), and (3) concepts that help divide the content into descriptions of 'Study' and 'Assay' according to the Reporting Structures for Biological Investigations (RSBI) working group recommendations for future modularization of checklists with a view to future integration [10, 11]; under 'Study' sit the concepts Organism, Phenotype, Environment, and Sample Processing and under 'Assay' falls the concept of Data Processing. This re-factoring into an 'Investigation' is a sign of the GSC's strong commitment to harmonizing our efforts with the rest of the 'omic standardization community. The MIGS checklist has been registered in the MICheck project and the GSC aims to be actively involved in the development of the MICheck community and strongly supports its goals [11].

The way in which genomes are described in our public databases has directly evolved from the way in which we describe even the shortest and simplest pieces of DNA sequences without special attention to information such as the geographical origin of the sequence. Significant efforts are underway by the INSDC to adapt and extend the infrastructure for describing genomes through the Genome Projects initiative [12]. The INSDC efforts are open to evolution, albeit at a conservative pace [12], and we would ideally like to see the entire MIGS specification within the Genome Projects initiative. The current checklist is presented in Table 1 [PLACE ON WEB?] and a mapping to the INSDC feature table is provided. Fields which are not already represented in INSDC can be placed into INSDC documents using the CC block (comments) or as a /note qualifier within the source feature [12].

The Genome Catalogue – a community resource

The issuance of a checklist must be further supported by an appropriate reporting structure (file format) for capturing data, a data exchange format, software, databases, and the development of appropriate controlled vocabularies and/or ontologies for expressing the

terms used in the annotations. The GSC is working towards these combined goals. We have implemented the MIGS checklist as an XML schema (migs.xsd) and built a database system that can generate customized forms automatically, and 'on-the-fly' from the schema for the sake of data input. The genome Catalogue (GCat) allows users to generate MIGS compliant genome reports through web forms as well as view and search existing genome descriptions through the online interface. Any changes to the underlying schema are immediately recognized and translated into changes in all the relevant parts of the GCat instance.

The GSC is also working in the area of ontology development, primarily through interactions with the Functional Genomic Investigations Ontology (FuGO) [13] and the collation of controlled vocabularies already in use in the community. GCat supports the use of controlled vocabulary terms and the capture of new terms for vetting by an appropriate authority. When terms are used for the first time they are given the status of 'proposed' and a user must provide an accompanying definition and source. Once approved, the terms are marked 'approved' (or pending, if awaiting updates or resolution of any conflicts).

The primary goal of GCat is to aid in the rapid prototyping and implementation of checklists. This system is generic and could be applied to the capture of more expressive sets of metadata from subsets of genomes. As long as a given checklist can be rendered into an XML schema, the GCat system can be used to implement an online data capture system. GCat is built XML technologies which are w3c recommendations and the beta version source code is available from gensc.sf.net.

Increasing the transparency of genomic databases

As described above, calculated information derived from genomic sequences is often subject to frequent change and therefore should be acquired directly from those conducting individual analyses. An ever increasing number of databases containing genomic annotation and other analyses are appearing, but more could be done to improve the transparency of the information contained in these resources. For example, the GenomeMine [5] and the GenomeAtlas [14] databases both support downloads of all stored datasets (e.g. as spreadsheets). The developers of the GenomeMine hope in the future to develop the Genomic Metadata Exchange Format (GnoME) which captures the provenance of the dataset and definitions for all the variables it contains. The issue of making genomic annotations more widely accessible for the sake of comparison and integration could be addressed through the use of the General Feature Format (GFF3) (http://song.sourceforge.net/gff3.shtml). There are numerous tools that support the reformatting of a variety of file types into GFF3, so generation of appropriate files is simple. The availability of a wide suite of tools for downstream analyses for all public genomes packaged in GFF3 format also means that users could combine the weight of evidence from many sources when examining a particular genome. This could reveal instances of systemic bias and therefore lead to better genomic annotations, as more composite features would be available and conflicting annotations could be highlighted for resolution. Combined with an approach like that employed in GnoME to capture provenance and if provided by each participating database through web services which would enable automatic harvesting of the data by other database providers, the circulation of GFF formatted datasets could revolutionize the transparency of content of the growing family of genomic databases.

The Future

The effort required to achieve the degree of 'transparency' advocated here is considerable but offers significant, obvious, and immediate benefits. The GSC has a standing open call for

participation, especially for the completion of case studies and genome reports that will help inform the definition of the MIGS specification and the collection of controlled vocabulary terms. The GSC plans to hold the 3rd and 4th GSC workshop at the National Institute for Environmental e-Science Centre (NIEeS) in Cambridge, United Kingdom in September 2006 and June 2007 to map out future activities.

Acknowledgements

We would like to thank NIEeS and the European Bioinformatics Institute (EBI) for hosting the first two GSC workshops and NERC for providing funds for co-ordination activities.

References

Table 1

Taxonomic Group	Organism	Phenotype	Environment	Sample Processing	Data Processing
Common	Complete taxonomic / genetic lineage information (below lowest rank recognized by NCBI taxonomy); Enough information to provide unencumbered access to genomic reagents (strain)		Latitude and longitude of sample; Time and day of collection; Depth / altitude; Habitat type; Description of environment;		Type of Sequencing method used; Estimated error rate and method of calculation
Eukaryotes	Is this a model organism; Number of chromosomes; Ploidy level; Estimated size; Reproductive mode (1)	Trophic level			
Prokaryotes	Reference for the description of the strain / sample; Information on whether	Growth conditions; Isolation conditions; Relationship to oxygen;	Environment (could be a host)		

	access to the isolate sequenced is restricted in any way; Identifiers for two culture collections	Relationship to host (pathogen etc); Presence of extrachromosomal elements; Reproductive strategy			
Plasmids	Host **(1)** Host range if known	Phenotype: Encoded traits like antibiotic resistance	Environment (medical, environmental, plant etc; same for hosts)		
Viruses	**(1)** Specific source of sample; Health/disease status of source host at time of collection	Whether normally pathogenic or not			
Organelles	**(1)** Full taxonomic information; Voucher condition and location				
Metagenomes	Expected number of organisms in the sample (community)			Sampling strategy (was it enriched, screened, normalized) volume of sample; justification for sampling methodology; process (e.g. how many clones)	

Table 1. The MIGS checklist. Fields which the community would like to see captured in addition to information already capture in INSDC genome annotation files and through the Genome Projects database. [NEED TO DISCUSS]

References

1. Overbeek, R. et al. *Nucleic Acids Res.* **33**, 5691-5702. (2005).

2. Field, D. et al. *Comparative and Functional Genomics* **6**, 357-362 (2006).
3. Field, D., Morrison, N., Sterk, P. & Selengut, J. *OMICS: A Journal of Integrative Biology* **(in press)** (2006).
4. Field, D. & Hughes, J. *Microbiology* **151**, 1016-1019 (2005).
5. Martiny, J.B.H. & Field, D. *Ecology Letters.* **8**, 1334-1345 (2006).
6. Ussery, D.W. & Hallin, P.F. *Microbiology* **150**, 2015-2017 (2004).
7. Haft, D.H., Selengut, J.D., Brinkac, L.M., Zafar, N. & White, O. *Bioinformatics* **21**, 293-306 (2005).
8. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. *Nat Rev Genet.* **5**, 335-344. (2004).
9. Garrity, G.M. (ed.) Bergey's Manual of Systematic Bacteriology, Vol. 1, 2nd. Ed. (Springer-Verlag, New York; 2001).
10. Sansone, S.-A., Rocca-Serra, P., Tong, W., Fostel, J. & Morrison, N. *OMICS: A Journal of Integrative Biology* **(in press)** (2006).
11. Taylor, C. et al. *Nat Biotechnol* **(submitted)** (2006).
12. Morrison, N. et al. *OMICS: A Journal of Integrative Biology* **(in press)** (2006).
13. Whetzel, P.L. et al. *OMICS: A Journal of Integrative Biology* **(in press)** (2006).
14. Hallin, P.F. & Ussery, D.W. *Bioinformatics* (2004).